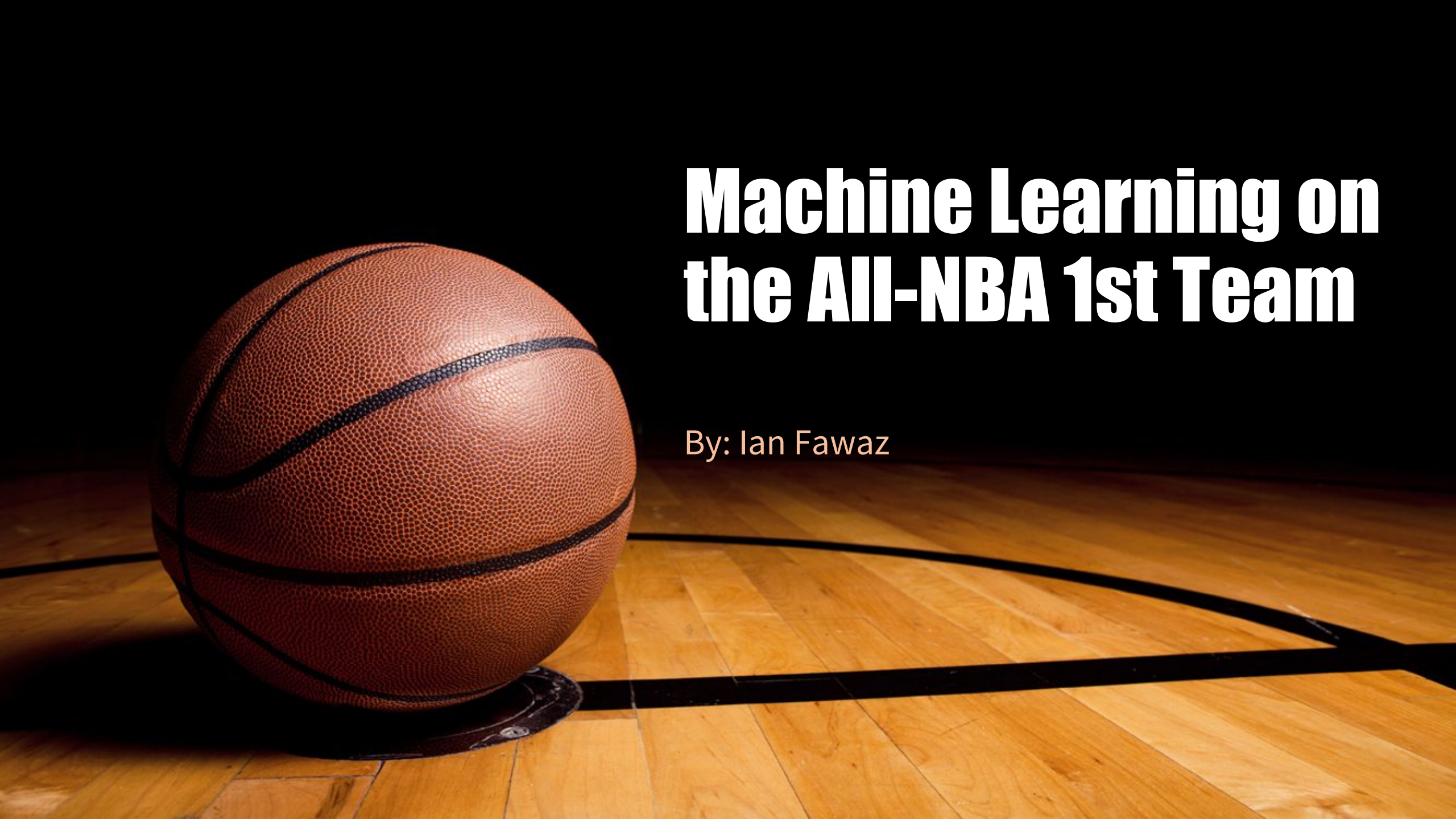


Machine Learning on the All-NBA 1st Team

By: Ian Fawaz

What is the All-NBA 1st Team?

- Voted on by panel of 100 sports media members at end of season
- The two best guards, two best forwards, and best center to create the top five man lineup

Contents of Dataset

- Every single NBA players' season stats dating back to 1950
- 47 relevant statistical floating point features
  - 'Year','Age','G','GS','MP','PER','TS%','3PAr','FTr','ORB%','DRB%','TRB%','AST%','STL%','BLK%','TOV%','USG%','OWS','DWS','WS','WS/48','OBPM','DBPM','BPM','VORP','FG','FGA','FG%','3P','3PA','3P%','2P','2PA','2P%','eFG%','FT','FTA','FT%','ORB','DRB','TRB','AST','STL','BLK','TOV','PF','PTS'
  - Will get into what some of these mean in a later slide
- Only used data from 1982 and beyond
  - In earlier years many features were blank in the dataset
  - 3-point line instituted in 1979, some features like 3-point attempt rate (3PAr) and amount of games started (GS) were blank until 1982 in the dataset

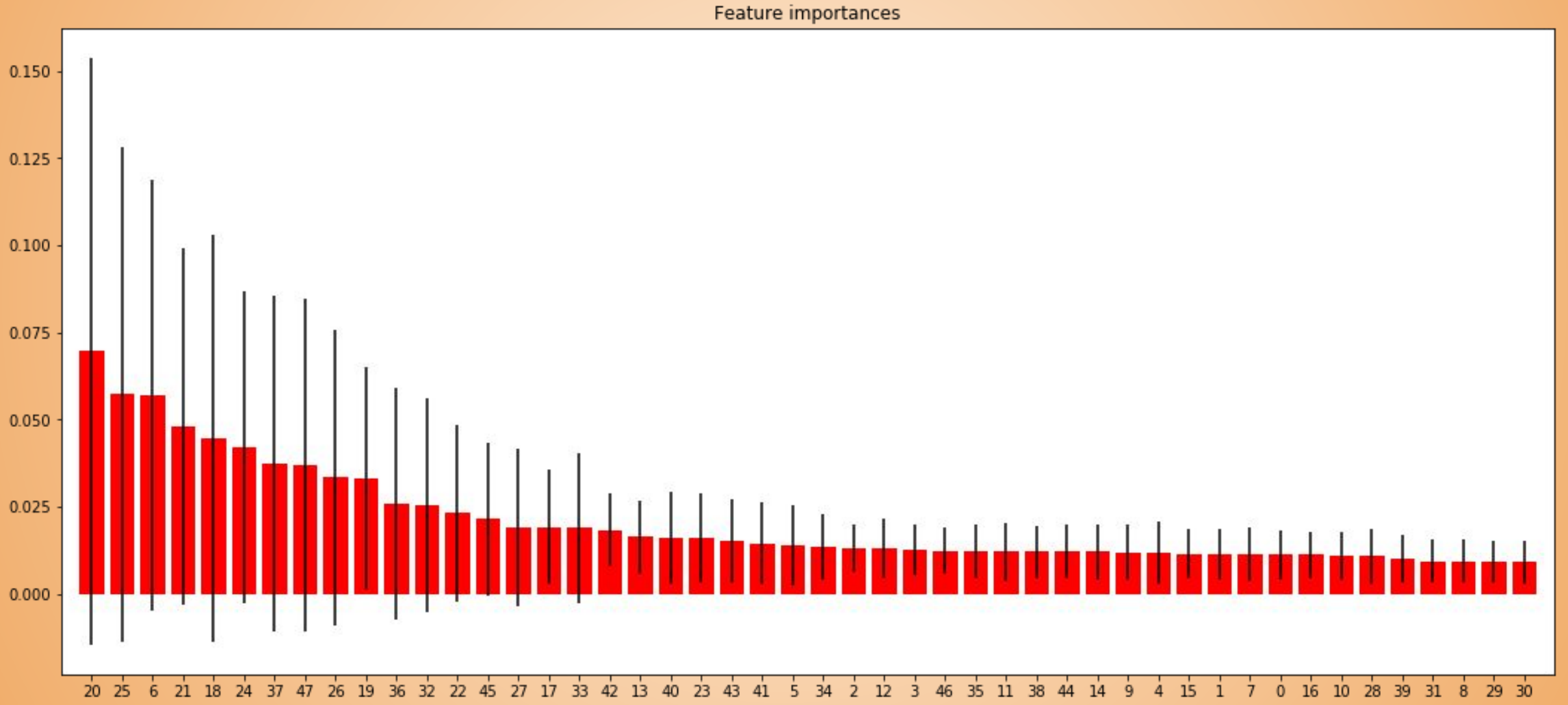
Contents of Dataset

- During classification, only used data from 1982 and beyond
  - In earlier years many features were blank in the dataset
  - 3-point line instituted in 1979, some features like 3-point attempt rate (3PAr) and amount of games started (GS) were blank until 1982 in the dataset
- Data Preparation
  - Created binary label column called “firstTeam” where all players were marked as ‘0’ unless they were on the all-NBA 1st team that season, in which case they were manually marked as ‘1’.
    - Individually marked 175 different player season rows as 1’s.

K-Fold Precision Results of Classifiers

- 10 folds, test_size was 15% of data
  - Random Forest average precision score: 0.8103
  - Naive Bayes average precision score: 0.1273
  - K-Nearest Neighbors average precision score: 0.7402
  - Logistic Regression average precision score: 0.7870

Error Analysis: Feature Importances



Error Analysis: Feature Importances

Top 10 features:

1. WS (Win Shares, encoded as ‘20’)
   - Metric to evaluate how many “wins” a player contributes to his team’s win total
2. VORP (Value Over Replacement Player, encoded as ‘25’)
   - Number of points a player is generating over a replacement player (average player).
3. PER (Player Efficiency Rating, encoded as ‘6’)
   - Takes into account both positive and negative player statistics at a per minute basis
4. WS/48 (Win Shares per 48 minutes, encoded as ‘21’)
5. OWS (Offensive Win Shares, encoded as ‘18’)
6. BPM (Box Plus/Minus, encoded as ‘24’)
   - Per 100 possession stat, evaluates player box score performance in relation to team performance and in comparison to league average
7. FTA (Free Throw Attempts, encoded as ‘37’)
8. PTS (Total Points encoded as ‘47’)
9. FG (Total Field Goals Made, encoded as ‘26’)
10. DWS (Defensive Win Shares, encoded as ‘19’)

Script source: http://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html

Feature Selection Experimentation

- Deleted all features except for the top 10
- Ran the Random Forest Classifier on this new feature set
  - Got very similar precision score of 0.8133, was 0.8103 with original feature set

Changing Hyperparameter in Random Forests

- Previously have set n_estimators, or the number of trees in the forest, at 500.
  - Maybe can reduce variance and overfitting by reducing the number of trees
  - Will change the random_state as well just to change things up

Changing Hyperparameter in Random Forests

- With original feature set, we run Random Forest several times with a different value each time for n_estimators, observing the K-Fold average precision scores
  - n_estimators = 5
    - Average precision score: 0.7373
  - n_estimators = 10
    - Average precision score: 0.8074
  - n_estimators = 50
    - Average precision score: 0.8293
  - n_estimators = 100
    - Average precision score: 0.8121
  - n_estimators = 250
    - Average precision score: 0.8221
  - n_estimators = 500
    - Average precision score: 0.8170
  - n_estimators = 1000
    - Average precision score: 0.8180
  - n_estimators = 10000
    - Average precision score: 0.8214

Conclusions

- May be better to use 50 trees for the Random Trees classifier on this dataset, reduces variance and perhaps overfitting
- Reducing the amount of features provided negligible difference
- Advanced metrics like Win Shares, VORP, PER, and BPM all do a good job of recognizing player value and are much more effective than simply looking at basic traditional metrics like PPG.
- Future applications?
  - Possibly testing on season stats at end of season to predict all-NBA 1st team
    - Would probably require some further tuning and looking into individual misclassifications
    - Another problem is the two guards, two forwards, and one center rule